École Polytechnique Fédérale de Lausanne

Analysing Public Opinion about Technology during the Second Industrial Revolution

by Emmanuelle Denove

Master Semester Project

Prof. Jérôme Baudry
Project Advisor

Dr. Elena Fernández Fernández
Project Supervisor

EPFL CDH DHI LHST
INN 139 (Bâtiment INN)
Station 14
CH-1015 Lausanne

June 9, 2023

Abstract

This project aims to analyse public opinion on technology during the Second Industrial Revolution. More specifically, it uses computational methods on newspaper articles from four different newspapers based in Germany, France, the United States and Spain to detect topics in different decades and journals that can later be compared. These topics are detected using Topic Modeling. They are then transformed into word embeddings in a multi-dimensional space using cross-lingual word embeddings to tackle the multilingual aspect. Finally, the topics are connected into a network in which clusters are detected. The topics within these clusters as well as their similarities and deviations across time and countries can help us interpret both how each country affronted emerging technologies, and how their approach compared to other countries.

Acknowledgements

I have thoroughly enjoyed working on this project with Elena, who always helped and supported me throughout the semester and allowed me to learn many new things on History and Cultural Studies that I didn't know before. I would also like to thank Jérôme, who was always available for any questions I had.

Finally, I would like to thank my family and friends, who were always available for a break by the lockers when needed, and the *Studio Ghibli Compilation* playlist on Spotify which has accompanied me throughout this project.

Contents

Introduction 1

This semester project is embedded into the Marie Curie Post-Doctoral grant lead by Dr. Elena Fernández Fernández (Principal Investigator): Time, Technology, and Globalization. A study of the role of technology in processes of modernization and globalization using the Press, Big Data, and Computational Research Methodologies (GLOTECH), grant agreement No 101024996.

1.1 Goal of the Project

The goal of this project is to analyse public opinion about technology during the Second Industrial Revolution by performing topic modeling on different countries' newspaper articles from that time.

The term "public opinion" is quite vague. According to Jürgen Habermas [1], it is a concept that exists as the result of the "public sphere", which he defines as the ensemble of "every conversation in which private individuals assemble to form a public body [...] in an unrestricted fashion". It therefore relies on both freedom of speech and freedom to assemble. This makes it a novel concept in the 19th century, when civil rights were first introduced in Europe and a country's inhabitants became its citizens. These citizens additionally need to have access to information in order to reliably discuss matters of general interest ; as such, Habermas defines newspapers as "the media of the public sphere". Public opinion naturally follows from the public sphere as "the tasks of criticism and control which a public body of citizens [...] practices vis-à-vis the ruling structure". When the bourgeoisie emerged as a new social class that had power but were private individuals separate from the state, they quickly took hold of newspapers which they used "against the public authority itself", turning them into an expression of public opinion.

In relation to this project, by analysing how and when different technologies were being discussed, or not discussed, in newspaper articles, we can get an idea of what the public thought of these technologies as they were emerging and developing.

We must also define the time frame in which to situate this analysis. Although there is some debate around it, the Second Industrial Revolution is usually defined from 1870 to 1914 [2]. However, some of the data collected for this project dates back to the 1820s and as far as the 1940s ; all of this data will be included in the analysis.

1.2 Historical Context

The 19th century was a period marked by unprecedented changes in the Western world. The combination of technological developments and societal upheaval completely revolutionised the world as it was known

in the late 1700s. Eric Hobsbawm describes this phenomenon in The Age of Revolution [3] and The Age of Capital [4].

On one hand, a new social class was emerging with the bourgeoisie, who had earned large amounts of money, for the first time being able to compete with the aristocracy. This increase in income was due both to new technologies eliminating certain costs of production, and to the cotton industry in the colonies which was generating immense profits. Society was becoming capitalist ; it believed that economic growth would be assured through competitive private enterprise.

On the other hand, the French Revolution had introduced the concept of citizenship and liberty of the people, launching a series of revolutions around 1848 meant to overthrow the aristocratic rulers in many European countries. These revolutions did not last long, and the old orders were quickly restored, but they had sufficed to instill enough fear in the aristocratic class to make some concessions of their power.

These revolutions complemented and enforced each other, the first creating a new potential governing power with the bourgeoisie, the second discrediting the traditional rulers. According to Eric Hobsbawm, "it was accepted that [. . .] money governed". Slowly, society stopped being dominated by agrarian cycles of the four seasons, and started being dominated by trade cycles and the rising and falling of the stock market.

At the same time, the development of new, revolutionary fast transport methods were contributing to a world that was becoming closer than ever. Our analysis situates itself in this time frame, when the world had been rapidly changing for decades and the happenings in one country were starting to have an impact in others.

1.3 Overview of Methods

The methods used in this project are the following :

- extraction of the raw data from .xml, .json and .txt files
- filtering of the raw data by certain keywords referring to certain technologies
- if needed, splitting the filtered data into individual newspaper articles with one coherent topic
- using the Pachinko Allocation Model to detect topics in these articles across newspapers and time
- determining word embeddings for the vocabulary to represent words by their semantic closeness
- using MUSE's cross-lingual word embeddings to map word embeddings from all languages into a single multidimensional space
- defining a distance between topics, represented as a set of words from the vocabulary, with the Word Mover's Distance
- creating a network of related topics across different newspapers and different time spans
- detecting communities of topics within this network

In the following sections, we will describe these methods in detail and report the results.

Data 2

2.1 Corpus presentation

The corpus contains four newspapers from four different countries :

- **Le Figaro** : this French daily newspaper based in Paris was founded in 1826 and exists to this day. Whilst originally meant to comment on the arts in a comical and satirical way, it had developed into a political newspaper by the middle of the 19th century. It became very popular quite quickly, and finally became France's leading daily in the late 1930's [5]. The data in the corpus is open-source, provided by the Bibliothèque Nationale Française. It contains 28502 documents in .json format, each representing one entire newspaper from a given day between January 1st, 1826 and November 24th, 1942.

Figure 2.1: First page of the first issue of *Le Figaro*, from January 1st 1826
https://gallica.bnf.fr

- **El Imparcial** : this Spanish daily newspaper based in Madrid was in circulation from 1867 to 1933. It had a liberal ideology and was one of the first newspapers in the country to belong to a private company as opposed to a political party. By 1890, it was one of Spain's leading journals.[6]. The data in the corpus is open-source, provided by the Biblioteca Nacional de España (BNE). It contains 23008 documents in .txt format, each representing one entire newspaper from a given day between January 1st, 1867 and May 30th, 1933.

- **Neue Hamburger Zeitung** (NHZ) : this German daily newspaper based in Hamburg existed between April 1896 and August 1922, when it was merged with the **General-Anzeiger für Hamburg-Altona** (founded in 1888) to simply become the *Hamburger Anzeiger*. These newspapers were both non-partisan, albeit left-leaning, and addressed respectively the bourgeoisie and the working class. The *Hamburger Anzeiger* existed until 1945, when it was merged once more with other Hamburg-based journals [7]. The data in the corpus was provided by email from Kerstin Wendt from the archives of the Staats- und Universitaetsbibliothek Hamburg Carl von Ossietzky. It contains 379671 documents from these two newspapers and their joint successor in .xml format, each representing one article from issues between September 2nd, 1888 and May 2nd, 1945.

- **New York Herald** (NYH) : this American daily newspaper based in New York was founded in 1835 and was in circulation until 1924, when it was bought up by its rival. It characterised itself by its explicitly politically neutral position. It was immensely popular : by 1845, it had become the most profitable daily in the United States [8]. The data in the corpus is open-source, provided by the Library of Congress. It contains 147709 documents from this newspaper in .txt and .xml format, each representing one page of an issue from a given day between January 1st, 1836 and December 31st, 1922.

Table 2.1: Summary of the corpus data

	Le Figaro	El Imparcial	NHZ	NYH
Corpus Begin	1826	1867	1888	1836
Corpus End	1942	1933	1945	1922
Num. of docs	28502	23008	379671	147709
Format	.json	.txt	.xml	.json/.xml

These files are originally compressed and distributed into a large number of folders. In order to load them into a .csv format, we first extract the archive, then iterate over all files in the resulting directory. For each file, depending on its format, the content of the document is loaded either by reading the lines of the file (.txt) or extracting the appropriate field (.json/.xml).

These newspaper data sets don't have a consistent number of articles over time. We plot the distribution of the number of documents by newspaper by year in Figure 2.2.

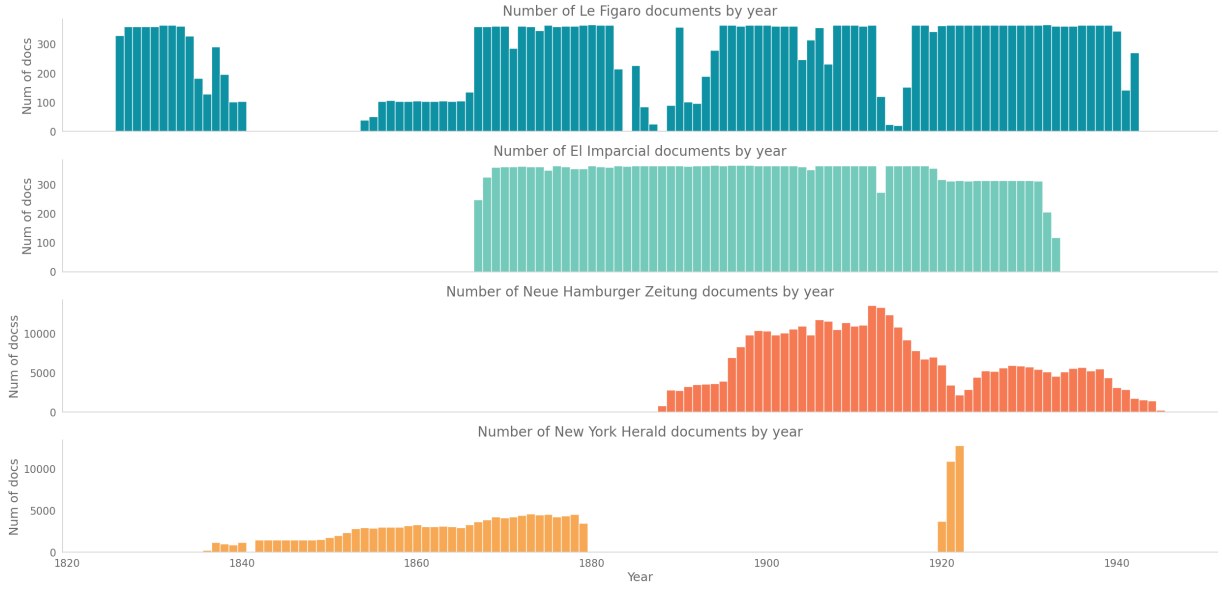


Figure 2.2: Distribution of the number of documents by newspaper by year

2.2 Corpus quality

The raw data for these newspapers was gathered by digitising the original issues using Optical Character Recognition (OCR). Since this method is not reliable and regularly results in errors, determining the corpus quality is necessary. To achieve this, Thomas Bench's OCR-qualification pipeline [9] was used. It represents text quality as the proportion of words in a document that were correctly digitised, which is determined by checking whether they exist in the respective language's dictionary. Assuming that words are either digitised correctly or not (i.e. no word is digitised as a different valid word), this metric gives us a percentage of overall text correctness. The dictionaries used here are the ones provided by the Enchant spellchecking library.

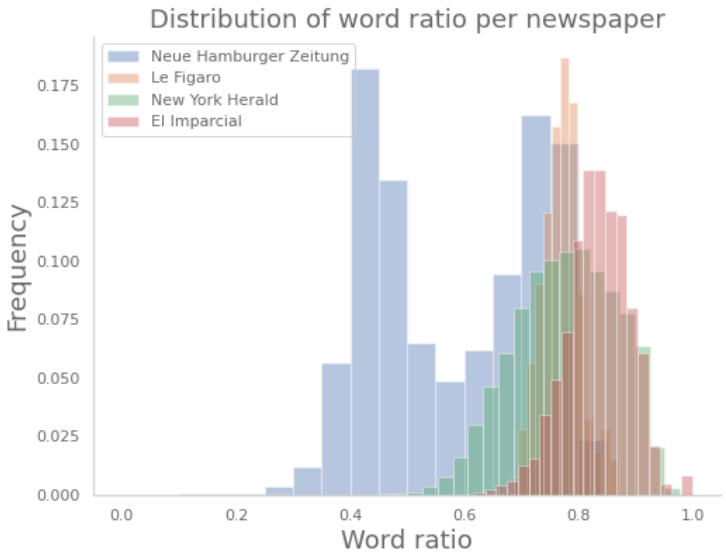


Figure 2.3: Distribution of word ratio per newspaper

The above Figure 2.3 shows that the articles in *El Imparcial* have the best quality, closely followed by the *New York Herald* and *Le Figaro*. The *Neue Hamburger Zeitung* performs far worse.

Table 2.2: Average word ratio of the OCR for each newspaper (LF = Le Figaro, EI = El Imparcial, NHZ = Neue Hamburger Zeitung, NYH = New York Herald)

LF	EI	NHZ	NYH
76%	83%	58%	77%

.

2.3 Article Splitting

As mentioned in section 2.1, the data of the different newspapers was gathered into "documents" in different ways, namely :

- **Le Figaro** : one "document" is an entire newspaper for a given day, with no clear delimitation between articles.
- **El Imparcial** : like for *Le Figaro*, one "document" is an entire newspaper for a given day. However, very often, different articles are separated by a newline.
- **Neue Hamburger Zeitung** : there is one "document" per article.
- **New York Herald** : one "document" represents one page of a newspaper issue for a given day. Within these pages, there is no clear delimitation between articles.

For topic modeling, this lack of delimitation between articles is problematic. Since entire newspaper issues, or even a single page of one newspaper issue, often contain many articles that tackle different topics, trying to perform topic modeling on these unaltered documents would lead to very confused and imprecise results. Therefore, a pipeline to split these documents into individual articles is necessary. These pipelines were determined empirically.

2.3.1 El Imparcial

Since most articles in the *El Imparcial* dataset are split by a newline, we simply consider every line of each document of length in characters $l > 20$ a distinct article. This method was adopted from Elisa Michelet's master project [10].

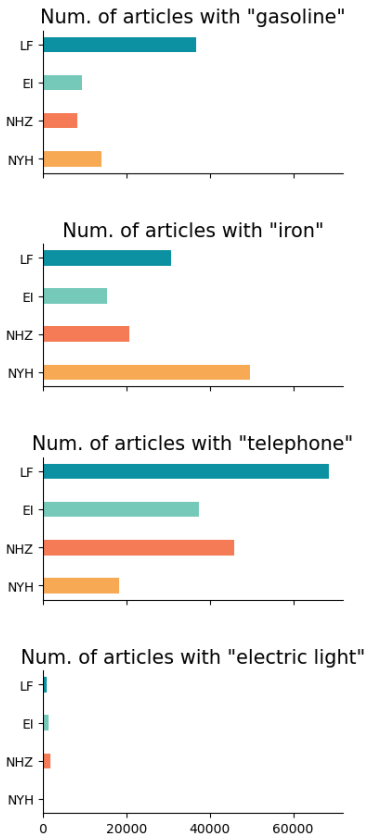


Figure 2.4: Number of models for each keyword for all newspapers (LF = *Le Figaro*, EI = *El Imparcial*, NHZ = *Neue Hamburger Zeitung*, NYH = *New York Herald*)

2.3.2 Le Figaro

The documents in this dataset contain many newlines that don't necessarily denote new articles. However, new articles are preceded by a newline, and many of their titles are written in uppercase. We therefore consider the beginning of a new article any newline followed by a string of length l whose alphabetic characters are uppercase with a frequency larger than k. Empirically, we determined for best results $l = 30$ and $k = \frac{2}{3}$.

2.3.3 The New York Herald

This dataset provided the .xml documents from the OCR as well as their textual interpretation. The XML files contain information about the size in pixels of the detected words. Since most titles of articles are written in a larger font than the rest of the text, this allows us to split articles based on this size information. More specifically, we consider the beginning of a new article any part of the text such that there is a sudden change in font size $diff_{size} < f$. Additionally, we require that the first n words of an article (i.e. the title) have uppercase characters with a frequency larger than k. Empirically, we determined $f = -20$, $n = 2$ and $k = \frac{2}{3}$.

2.4 Keyword Detection

To determine the conversations surrounding different technologies in the newspapers of our corpus, the articles were filtered and only those that include the given keyword were kept.

- **Electric Light** : contrarily to the other three words, this keyword is made up of two distinct words. To correctly filter the articles that contain this keyword, only those that included both words within $n = 100$ characters of each other were kept.
- **Gasoline** : since this word has a synonym, namely **petrol**, articles containing any of these two words were considered. The French translation of gasoline, *"essence"*, also has the meaning of the English word "essence", which resulted in many false positives when filtering that dataset.
- **Iron** : the distribution of number of articles including *"iron"* by newspaper can be seen in Figure 2.4.
- **Telephone** : the distribution of number of articles including *"telephone"* by newspaper can be seen in Figure 2.4.

The exact words used to filter are summarized below. Since "electric light" yielded so few articles, it is discarded from future analysis.

Table 2.3: Keywords used to filter articles

	Electric Light	Gasoline	Iron	Telephone
French	lumiere AND electrique	essence OR petrol	" fer "	telephone
Spanish	luz AND electrica	gasolina OR petroleo	" hierro "	telefono
German	elektrisch AND licht	benzin OR kraftstoff	" eisen "	telefon OR telephon
English	electric AND light	gasoline OR petrol	" iron "	telephone

Topic Modeling 3

The goal of this project is to detect the nature and evolution of discussions surrounding technology in different countries over time. In order to determine how these subjects were discussed in the newspapers in our corpus, we use topic modeling. This is a widely-utilized generative probabilistic method that can be used on large sets of unclassified text to detect clusters of words that often occur together, referred to as "topics". It is able to both capture synonyms and differentiate instances of words that have multiple meanings [11]. It also works independently of the language of the text, making it well-suited to the multilingual aspect of this project.
The methods used here were adapted from Elisa Michelet's Master Project [10] and Germans Savcisens' pipeline for Pachinko Allocation [12].

3.1 Pachinko Allocation Model

The Pachinko Allocation Model (PAM) was used to detect topics in documents. This model is built on the Latent Dirichlet Allocation (LDA) method, which uses a Dirichlet distribution to model documents in a corpus or words in a vocabulary as a distribution over a finite set of topics [13]. This distribution can then essentially describe the contents of the document, or the topics to which a particular word is referring to.

The Pachinko Allocation Model [14] extends on the LDA by additionally capturing correlations between topics. It models the vocabulary as the leaves of a Directed Acyclic Graph (DAG), and the topics as the nodes. Correlations are then represented by the edges between different nodes and leaves of the graph, which can be between words or between topics, the latter resulting in a set of super-topics. A classic 4-level PAM consists of a root, a set of super-topics, a set of sub-topics and a vocabulary.

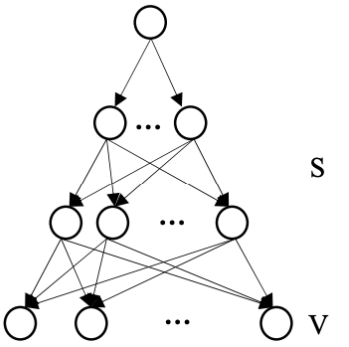


Figure 3.1: The DAG in a four-level PAM

Since PAM performs much better when the number of super-topics and sub-topics is pre-defined [15], a grid search is performed to obtain the optimal parameters. This process is detailed below.

3.2 Best parameter calculation

To train a Pachinko model, we require two arguments k_1 and k_2 denoting respectively the number of super-topics and the number of sub-topics in the corpus . Since we do not know these beforehand, we create the model with all possible pairs of k_1 and k_2 with $k_1 \in [1, 2]$ and $k_2 \in [k_1, 14]$, resulting in 27 different models. These values were chosen empirically.

To determine the best parameters, we compare their coherence value, which is a metric that aims to quantify the coherence and human interpretability of the top words per topic returned by a topic model [16].

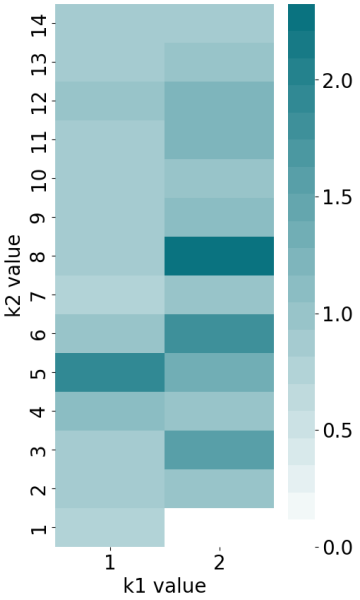


Figure 3.2: Mean Test Score for all tested

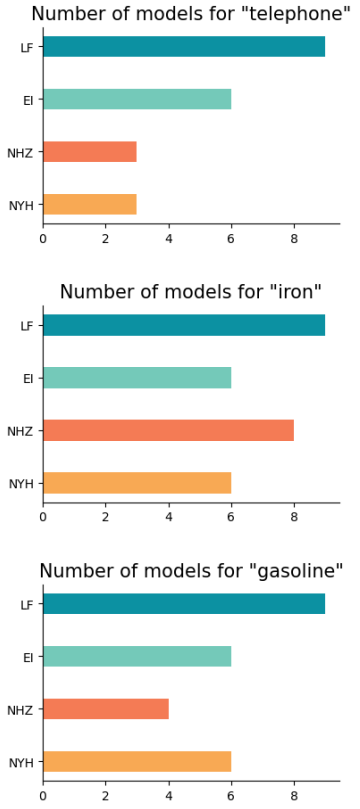


Figure 3.3: Number of models for each keyword for all newspapers (LF = *Le Figaro*, EI = *El Imparcial*, NHZ = *Neue Hamburger Zeitung*, NYH = *New York Herald*)

Specifically, the chosen metric is c_V-coherence because its results are the closest to a human interpretation [10].

By default, the method used to determine the best parameter is the one-standard-error rule, which selects the model whose prediction error is at most one standard error worse than that of the best model [17]. However, this method is often quite ungenerous and only detects 1 or 2 subtopics. In those cases, we used the parameters of the best model as determined by the grid search.

Since the optimal number of topics in a corpus changes based on the documents used to train the model, this calculation was repeated for every combination of keyword, newspaper and time-span. As an example, we report in Figure 3.3 the results of the mean test score for the grid search across all parameters for newspaper articles in the *Neue Hamburger Zeitung* containing the word *"Telefon"* ("telephone") between the years 1920 and 1935 (exclusive). The best test score was achieved by the values $k_1 = 2$ and $k_2 = 8$. However, the adopted parameters were the ones determined by the one-standard-error rule, namely $k_1 = 1$ and $k_2 = 5$ or 1 super-topic and 5 sub-topics.

3.3 Model training

Following the method used by Elisa Michelet [10], separate models were trained for every combination of keyword, newspaper and time-span in order to track the evolution of topics across time and newpapers ; this gave a total of 75 models. The final pipeline from raw documents to trained model for a given newspaper is:

- splitting the raw documents into articles to include the given keyword as described in section 2.3
- splitting the set of articles into time-spans. The default time-span of a split is 10 years, which was adjusted to be shorter if there were too many documents, or longer if there were too few, to improve the topic modeling
- using Spacy [18] to retrieve the lemmas of words in the articles, keeping only lemmas that are not stop words and that are alphanumerical. The lemma of a word refers to its basic form without grammatical modifiers (i.e. the lemma of "words" is "word" and the lemma of "has" is "have"). Lemmas that have less than 4 characters were also discarded, since they often represent noise from the OCR.
- determining the optimal number of super-topics and sub-topics for the given model as described in section 3.2
- using Tomotopy's Pachinko Allocation training model [19] to train the given model following the pipeline proposed by [12] with the previously determined parameters

Finally, each given model was represented as the set of its $k = 15$ most representative words as determined by the topic modeling. As an example, these are the 15 most representative words of one of the 13 topics detected in *New York Herald* articles containing the word *"iron"* from 1935 to 1950 : `['cure', 'disease', 'stricture', 'medicine', 'remedy', 'tooth', 'treatment', 'bottle', 'medical', 'pain', 'complaint', 'Pills', 'physician', 'consult', 'patient']`.

Cross-lingual Word Embeddings 4

After building a list of top $n = 15$ topics for each combination of keyword, newspaper and time span, we want to compare these topics across newspapers to determine the differences and similarities in discussion in different countries. This pipeline was adopted from Elisa Michelet's Master project [10].

4.1 Word embeddings

Our goal is to compare the topics in newspaper articles mentioning certain technologies over time and across countries. However, these topics are represented only as a collection of words that may have a different morphology despite having a similar meaning or being related in context. In order to compare them, we need a measure of their semantic similarity to each other. A popular method that achieves this is word embeddings, which maps words to vectors in a multi-dimensional space in which similarity can be measured mathematically [20].

4.2 Choice of Algorithms

There are different frameworks to perform these word embeddings. Following the evaluation done by Elisa Michelet [10], the chosen model is **Fasttext** [21]. This model builds on the word2vec skip-gram model proposed by Mikolov et. al. in 2013 [22]. The skip-gram model uses a shallow neural network with 3 layers, as shown in Figure 4.1. It takes as input a target word and outputs a prediction of the N words before and after the target word. The hidden layer holds the weights that generate the word embeddings, which are adjusted during training. Fasttext improves on this by breaking input words into character n-grams to include subword information in the neural network.

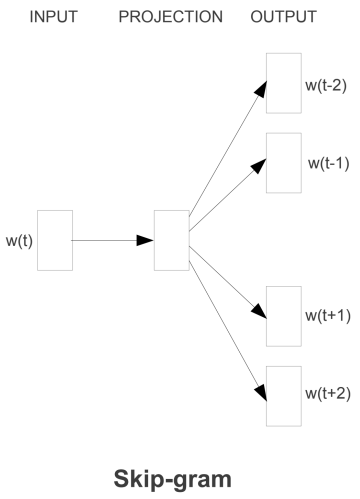


Figure 4.1: Sketch of the layers of the word2vec skip-gram model.

Since the datasets are so large, the Fasttext models are trained on the random samplings of 1000 articles for each newspaper and keyword, resulting in a total of 12 models.

4.3 MUSE cross-lingual embeddings

After creating word embeddings to compare semantically similar words within each newspaper, we want to do the same across the whole corpus. As such, the next step is determining **cross-lingual word embeddings**. This was done using unsupervised Wasserstein procrustes using the MUSE library [23], as proposed by Elisa Michelet [10].

5 Creating Topical Networks

Now that we have converted words of all languages into vectors that can be mapped onto a multi-dimensional space, we can detect clusters among the detected topics. These clusters can help us determine semantic closeness of topics detected in different newspapers at different times, allowing us to compare them. This pipeline is adopted from Elisa Michelet's Master project [10].

5.1 Word Mover's Distance

In order to cluster the topics, we first need to quantify the distance between them. For this, we use Word Mover's Distance, which calculates the distance between two documents as the minimum distance between the word embeddings of those documents [24]. By considering the list of top 15 words representing each topic as small documents, this metric can be used to determine the semantic "distance" between any two topics in our dataset.

More formally, consider $T = w_1, w_2, ..., w_n$ and $T' = w'_1, w'_2, ..., w'_n$ two topics defined by their top n words. Let $\mathbf{V} \in \mathbf{R}^{nxn}$ be a sparse flow matrix such that $\mathbf{V}_{w_i, w'_j} \geq 0$ denotes the distance traveled by $w_i \in \mathbf{T}$ to reach $w'_j \in \mathbf{T'}$. Then we define the Word Mover's Distance as :

$$wmd(T, T') = min_{\mathbf{V} \geq 0} \sum_{i,j=1}^{n} \mathbf{V}_{ij} \left\| w_i - w'_j \right\|_2$$

5.2 Network creation

The pipeline to create a network of semantically related topic clusters determined by Elisa Michelet [10] is as follows :

1. Set two thresholds $L_s > 0$ and $L_d > L_s$, respectively for topics of the same language and topics of different languages.
2. add a vertex V_{Ti} for every topic T_i
3. compute the pairwise distance between all topic pairs (T_i, T_j)
4. let $L = L_s$ if T_i and T_j are the same and $L = L_d$ otherwise
5. add an edge between V_{Ti} and V_{Tj} if and only if $wmd(T_i, T_j) \leq L$

5.3 Graph community detection

After creating the network of topics based on their semantic closeness, we want to detect clusters in this network which translate to clusters of topics from different newspapers and decades. This will allow for the comparison of topics across newspapers and time.

To create these graphs, two algorithms, both created with large networks in mind, were used based on what was more suited for each graph.

5.3.1 Clauset-Newman-Moore

The Clauset-Newman-Moore algorithm is a greedy algorithm that optimizes the modularity of a graph, meaning that its communities are strongly connected within each-other but weakly connected between each-other [25]. It measures this by calculating the deviation of the number of edges within a community for a given graph to its expected value in a randomised graph.

The algorithm is performed in two steps : first, each node is considered its own community. It then greedily joins communities pairwise to maximize the modularity and stops when that is no longer possible [25].

5.3.2 Louvain method

The Louvain method is an algorithm that focuses on modularity gain. It starts by considering each node as one community. It then tests whether modularity can be gained by moving a given node to a neighboring community ; if it can, it moves the node to the community that renders the highest gain, otherwise the node stays in its original community. It does this until it finds no more gain [26].

5.4 Best parameter determination

In order to find the optimal values for L_s and L_d as well as which of the aforementioned algorithms to use, we evaluate a range of values with both algorithms, specifically $L_s \in [0.5, 7[$ and $L_r \in [L_s + 0.25, 7[$ with a step of 0.25. All calculated communities are scored based on the following metrics :

- **Modularity M** ; we reuse the definition from Section 5.3.1
- **Performance P** quantifies the distinction of communities by summing the number of edges within communities and the number of non-edges between communities and computing its ratio with the total number of potential edges [10]
- **Coverage C** refers to the ratio of the number of edges within communities with the total number of edges in the graph [10].

The score is defined as follows :

$$score = \alpha M + \beta P + \gamma C$$

with $\alpha = 0.7$, $\beta = 0.15$ and $\gamma = 0.15$ determined empirically. The chosen parameters are then the ones that led to the best score.

5.5 Results

Table 5.1 shows the best parameters determined by the method above. After constructing graph communities with those thresholds, we obtain one graph per keyword clustering the topics of different newspapers and time spans. An example of such a graph is shown in Figure 5.1.

Table 5.1: Parameters for topic networks

	Gasoline	Iron	Telephone
L_s	6,25	4,75	2,25
L_r	6,75	6,75	5,75
algorithm	louvain	louvain	louvain

These networks practically only cluster topics that are from the same newspaper. Therefore, we can say that discussions surrounding these specific technologies were, in general, quite different across countries, at least in terms of frequently recurring semantics.

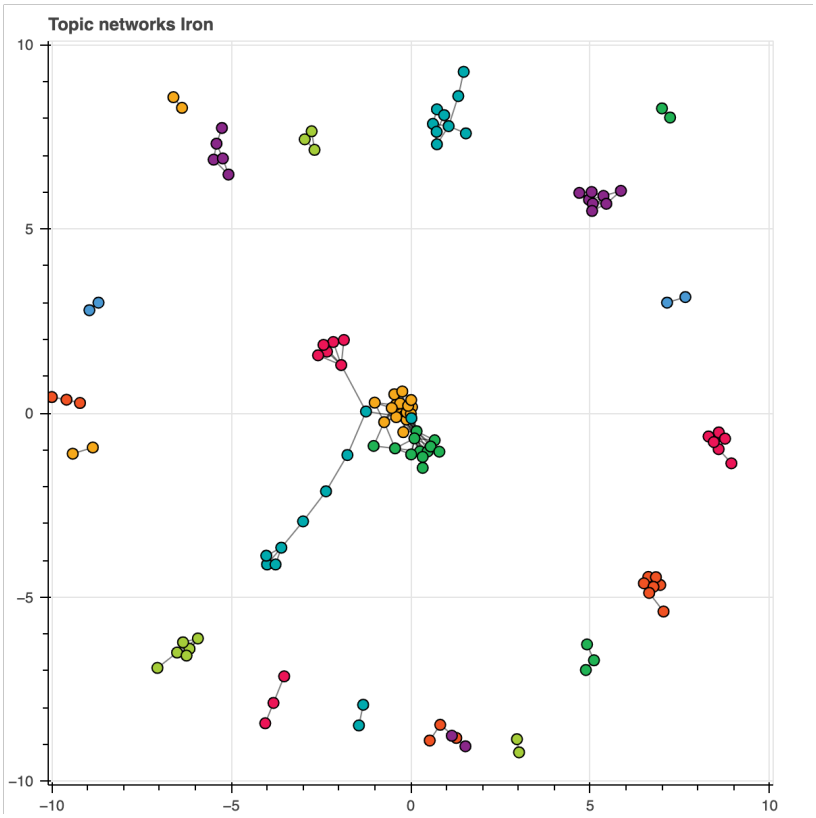


Figure 5.1: Topic Network for the keyword *Iron*

Results and Interpretation 6

This section discusses the findings from the topic modelling described in section 3 and the networks discussed in section 5 for each of the keywords *Gasoline*, *Iron* and *Telephone*. As mentioned in section 2.4, the keyword *Electric Light* was not further analysed due to the very limited number of articles that included it.

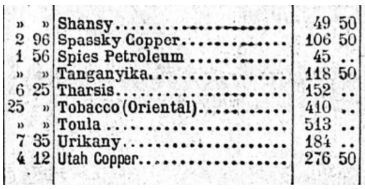
6.1 Gasoline

The term *"Gasoline"* first appears in the corpus around the second half of the 1860's and is mentioned in more and more articles as time goes on. It is consistently present in discussion around economy and the stock market, as well as transportation. However, the type of transportation it relates to changes over time as technology advances.

6.1.1 Economy

As is still the case today, the topic modeling shows that gasoline was a major factor in economic conversations around the turn of the century. In *Le Figaro*, it was listed in advertisments for local sales as early as the 1880s ; by the beginning of the 20th century, we can find the term very frequently in daily reports of stock market prices, as can be seen in Figure 6.1. Note that the word "petroleum" is in English and therefore referring to an Anglo-Saxon company, testifying to the beginning of an international stock market.
This pattern of finding a listing on gasoline in stock market reports carries over to the United States, where the words *"gasoline"* or *"petroleum"* were already featured almost daily in such lists in the late 1860'ss.
In the *Neue Hamburger Zeitung*, the economics of gasoline are often discussed in relation to companies beginning to use it to replace an older, more expensive technology. An article from the 10th of April 1912 reports on the proposal to start powering city cabs with gasoline, namely because "their maintenance is much cheaper than that of the electric car". Stock market movements are usually analysed in a long, essay-like format, as opposed to the systematic lists we find in the French and American newspapers.

In *El Imparcial*, the vast majority of mentions of *"gasoline"* is in relation to advertisements of some kind. The issue from the 26th of May 1889 promotes "Machines for industries. Tubeless petrol devices. Large number of specialties of notable utility". Similarly, Figure 6.2 shows an advertisement for a gasoline refinery. We also see this type of advertisement in the German newspaper, both for sales of gasoline-based products and for individuals looking to hire a mechanic.



»	»	Shansy	49 50
2	96	Spassky Copper	106 50
1	56	Spies Petroleum	45 ..
»	»	Tanganyika	118 50
6	25	Tharsis	152 ..
25	»	Tobacco (Oriental)	410 ..
»	»	Toula	513 ..
7	35	Urikany	184 ..
4	12	Utah Copper	276 50

Figure 6.1: Excerpt from an article from the 21st June 1911 issue of *Le Figaro*
https://gallica.bnf.fr

Figure 6.2: Advertisment for a gasoline company from the 10th December 1891 issue of *El Imparcial*
https://hemerotecadigital.bne.es/

6.1.2 Accidents

When it was first being used, it seems that there were frequent accidents due to gasoline. On the 28th of August 1870, the *New York Herald* reported "another explosion of gasoline". These reports can be found across all newspapers over the entire time-span. The *Neue Hamburger Zeitung* for example reported "22 people injured by an exploding petrol-tank" on the 8th of March 1929, 50 years later.

6.1.3 Boats

In the port city of New York, much of the discussion around gasoline is related to boats. Indeed, the *New York Herald* often dedicates large sections of every issue to "Shipping News", reporting on the comings-and-goings of ships and what they are carrying. This includes gasoline : a paper from 1877 mentions a ship whose "cargo consisted of 252.988 gallons refined petroleum" (19th Oct. 1977).

6.1.4 Car races


AUTOMOBILISME

LE RECORD DU KILOMÈTRE

Le défi lancé par M. Jenatzi à M. le comte de Chasseloup-Laubat pour la médaille d'or offerte par la *France automobile* au recordman du kilomètre se courra mardi prochain, 18 janvier, sur la route centrale du parc agricole d'Achères.

Figure 6.3: Announcement of a car race from the 14th January 1899 issue of *Le Figaro*
https://gallica.bnf.fr

Cars began being powered by gasoline within the time-span of the corpus, which makes them a recurring topic in articles, especially in the French corpus. Interestingly, the main subject related to cars, and as such to gasoline, is that of car races.
In *Le Figaro*, the sports column had a section on automobile races by the late 1890's. Figure 6.3 shows a snippet in that column announcing the next competition for the fastest driven kilometer. According to the text, this challenge was set by a Count. Indeed, from the discovered topic networks that cluster "car race" and "aristocracy", we can deduce that this was a sport reserved for the upper class.

6.1.5 Airplanes


EL VUELO ESPAÑA-FILIPINAS

Los capitanes Loriga y Gallarza, en Calcuta

Estévez y el mecánico Calvo han sido hospitalizados en Amman, donde se encuentran en estado satisfactorio.--Actos de simpatía y gratitud a la aviación británica

Gráfico de la etapa Agra-Calcuta, que los capitanes Loriga y Gallarza han cubierto en siete horas a una velocidad de 175 kilómetros, volando a una altura media de 2.000 metros

Figure 6.4: Excerpt from the 20th April 1926 issue of *El Imparcial*
https://hemerotecadigital.bne.es

The early 20th century saw the beginning of the gasoline-powered airplane, which was particularly commented by the German, French and Spanish newspapers. In *El Imparcial* and *Le Figaro*, a separate topic on airplanes emerges in the 1920's. In April 1926, Joaquín Loriga completed the first long-distance flight between Spain and the Philippines ; this event was covered on the front page of the newspaper (see Figure 6.4). In the *Neue Hamburger Zeitung*, the topic is detected 10 years later, often in relation to articles explaining in detail the problems encountered during a particular flight from the perspective of the pilot. The issue from the 21st January 1938 for example recounts an emergency landing on water after an engine failure.

6.1.6 Conclusion

From the above analysis, we can conclude that discussions around gasoline differ considerably across time and countries. In New York, it was seen mainly as part of the stock market, or as an imported and exported good in the city's port. In *Le Figaro*, it was also included in stock market reports, which were local at first but became international in the early 20th century. Once car races became popular in the 1890s, the term was also used in relation to that. In *El Imparcial*, "gasoline" is often the center term of advertisements. With the beginning of the gasoline-powered airplane in the 1920s, we can find many articles detailing the first flights and the problems they encountered in the French, German and Spanish corpus. The thoroughness of these articles testifies to the public's curiosity for this new method of transportation.

6.2 Iron

Iron was a major contributor to one of the biggest innovations of the 19th century : the railway. According to Eric Hobsbawm, "no innovation of the Industrial Revolution has fired the imagination as much as the railway. In *The Age of Revolution* [3], he describes how the invention of the railway changed the economic system into the capitalistic system we know today by providing immense investment opportunities for the newly rich. As such, the term *"iron"* is extremely prevalent in all newspapers across our corpus. Apart from trains, it is also discussed in contexts of the stock market, boats, household items and machinery.

6.2.1 Railways

In both German and French, the word for "iron" (respectively "Eisen" and "fer") is part of the word for railways (namely "Eisenbahn" and "chemin de fer") ; as a result, the articles filtered by "iron" in these languages are in the vast majority about railways. *Le Figaro* has been regularly announcing train departure hours and prices since the 1860s, as can be seen in Figure 6.5. In Germany, the *Neue Hamburger Zeitung* from the 31st July 1895 celebrates the 60-year anniversary of the first railway in Germany. In fact, the railways were so important in the Germanic area that Prussia had a railway minister around the turn of the century (see *Neue Hamburger Zeitung* from 14th November 1901).


COMPAGNIE DES CHEMINS DE FER
De Paris à Lyon et à la Méditerranée (Section nord)
SERVICE DIRECT DE :
PARIS A MILAN
Par Mâcon, Culoz, le mont Cenis, Turin, Vercell, Novare et Magenta
Durée du trajet : 39 heures par train express
DÉPARTS TOUS LES JOURS A 8 H. 40 SOIR
Billets valables pour 15 jours avec faculté d'arrêt à
Dijon, Mâcon, Culoz, Aix-les-Bains, Chambéry, Chamousset, Saint-Jean-de-Maurienne, Suze, Turin, Vercell (Palestro et la Sesia), Novare et Magenta.
Prix des places
DE PARIS A
AIX-LES-BAINS
CHAMBERY
CHAMOUSSET
TURIN
NOVARE
MILAN
1re classe 2e classe 3e classe
66 15 48 35 35 70
69 15 49 75 36 20
83 55 61 05 37 70
103 75 83 75 54 50
114 40 91 40 72 60
119 30 95 10 74 30
Correspondances : A Chamousset, pour Moutiers et Albertville (Diligence) ; à Saint-Jean-de-Maurienne, pour Modane et Lans-le-Bourg (Diligence) ; à Turin, pr Pignerol, Coni, Alexandrie, Montebello et Gênes (Ch. de fer) ; à Novare, pr Arona, (Sesto-Calende) et le lac Majeur ; à Milan, pr Bergame, Brescia, Monza, Camerlata, Comeet Varèse (Ch. de fer).
S'adresser, pour les renseignements, à l'administ. du chemin de fer Victor-Emmanuel, 48 bis, rue Basse-du-Rempart, et à la gare de Lyon, boul. Mazas, au bureau des correspondances, où sont délivrés les billets. Des voitures de poste, à 2, 3, 4, 5, 6 et 7 places, pour la traversée du mont Cenis, peuvent être retenues à ce bureau quelques jours à l'avance.

Figure 6.5: Excerpt from the 24th February 1861 issue of *Le Figaro*
https://gallica.bnf.fr

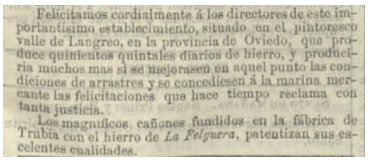
Some of this news about railways crossed national boundaries. In 1862, an article in *Le Figaro* comments that since the opening of the railway, "Hamburg has found itself at the centre of Europe, at only 14 hours from Brussels, 14 hours from Amsterdam [...] and 16 hours from Paris". Similarly, the *Neue Hamburger Zeitung* reported in 1894 that "from the construction of the East Chinese railway, no good news are coming in" (24th November 1894). The fact that news from Chinese railway construction not only reaches Germany but is considered of public interest shows the beginning of a world that is becoming closer and more interdependent.

6.2.2 Boats

Although the revolutionary form of transportation created using iron during this time-span was the railway, the material was still used to upgrade the more traditional ships and boats. The 11th August 1845 issue of the *New York Herald* marvels at the sheer size of the steamer Great Britain, which was on exhibition in New York, where "iron [is] being employed to a greater extent than in any other ship". A 1851 issue describes a company's purchase of the rights to manufacture "Montgomery's patent corrugated iron boilers" for their steamships (1851-06-17). Throughout the 19th century, the German and American newspapers, which are both based in ports, contain large sections detailing the use of iron in the construction of ships specifically. Interestingly, this conversation does not carry over to the landlocked cities.

6.2.3 War

During the period of the American civil war, the *New York Herald*'s regular descriptions of the features of new ships also extends to war ships, like "the rebel iron-clad vessel [...] mounting numerous guns of heavy calibre" (1862-08-26). Iron is also used in artillery : "twenty-four pound iron guns on [...] carriages" (*New York Herald*, 1862-03-08). During this period, almost 1/3 of articles containing "iron" are related to the war, which is a topic that does not appear in the previous or following time-spans.


Felicitamos cordialmente á los directores de este importantísimo establecimiento, situado en el pintoresco valle de Langreo, en la provincia de Oviedo, que produce quinientos quintales diarios de hierro, y produciria muchos mas si se mejorasen en aquel punto las condiciones de arrastres y se concediesen á la marina mercante las felicitaciones que hace tiempo reclama con tanta justicia.

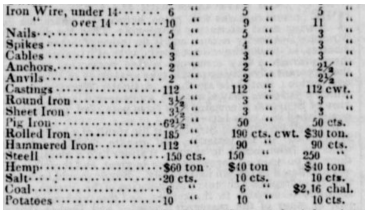
Los magníficos cañones fundidos en la fábrica de Trubia con el hierro de *La Felguera*, patentizan sus escelentes cualidades.

Figure 6.6: Excerpt from the 8th June 1867 issue of *El Imparcial* https://hemerotecadigital.bne.es/

In fact, technological innovation in the field of war machinery was celebrated : an article from 1867 of *El Imparcial* congratulates a Spanish iron factory that received a silver medal at that year's international exposition in Paris, iron which was mainly used to create "incredible canons" (see Figure 6.6).

During World War 1, discussions around iron became very centered on war in the European countries. A *Neue Hamburger Zeitung* article from 1912 states that "the German-English question can only be solved by blood and iron" (1912-08-23). An article from 1914 in the same newspaper reminds readers of the "important military meaning of railways" (1914-08-05) just as the war is beginning.

6.2.4 Economy



Iron Wire, under 11	6 "	5 "	5 "
" over 11	10 "	9 "	11 "
Nails	5 "	5 "	3 "
Spikes	4 "	4 "	3 "
Cables	3 "	3 "	3 "
Anchors	2 "	2 "	2½ "
Anvils	2 "	2 "	2½ "
Castings	112 "	112 "	112 cwt.
Round Iron	3½ "	3 "	3 "
Sheet Iron	3½ "	3 "	3 "
Pig Iron	62½ "	50 "	50 cts.
Rolled Iron	185 "	190 cts. cwt.	$30 ton.
Hammered Iron	112 "	90 "	90 cts.
Steell	150 cts.	150 "	250 "
Hemp	$60 ton	$10 ton	$10 ton
Salt	20 cts.	10 cts.	10 cts.
Coal	6 "	6 "	$2,16 chal.
Potatoes	10 "	10 "	10 cts.

Figure 6.7: Excerpt from the 7th August 1842 issue of the *New York Herald* https://chroniclingamerica.loc.gov/

Since a lot of people became incredibly rich thanks to the iron industry, it was constantly discussed in the context of the economy. In the *New York Herald*, we often find the term in listings of stock market prices and imported goods. Figure 6.7 from August 1842 shows a comparison of the duties to be paid on different types of iron imported from Great Britain after the introduction of a new bill.

It is also discussed in the context of descriptions of company production, testifying to its status as a major industrial good. A 1843 issue of the *New York Herald* describes the domestic industry of the state of Virginia, whose "chief manufactures are iron castings, bar iron, and glass" (1843-11-19).

As the world becomes closer, these discussions on economy do not stay local. In 1882, *El Imparcial* published an article discussing Spain's

commercial tires to Great Britain and used the importation of different types of iron to illustrate it (1882-11-08). Similarly, in 1921, a special correspondent of the *New York Herald* based in Berlin comments on the state of German industry after World War 1, namely that "by the transfer of Lorraine to France, Germany lost about three-fourths of its iron supply". Also in the context of the war, *Le Figaro* sums up reparations demanded of Austria in 1919, which include the obligation to deliver "during five years, wood, iron and magnesium", once again underlining iron as an essential material.

6.2.5 Infrastructure

Iron is also used for the construction of houses, and is advertised as such in real estate notices, like "two [..] brown font houses, with iron balconies, [...] renting for 500$ each" (*New York Herald*, 1856-02-25). A 1928 article from *El Imparcial* reports that the Madrid government, when deciding on the building material for a new viaduct, picked iron over stone because it was "more economical" albeit "less durable" (1928-10-25). This can be considered as an early case of prioritising cheapness over quality, which is considered very common nowadays.

When the international exposition was hosted in Paris in 1889, *Le Figaro* describes its "palace of machines" where "glass and iron combine to make the palace strong, elegant, brilliant and light". This choice of material for a building meant to represent progress and machinery shows how linked these concepts were in people's minds.

6.2.6 Housework and household items

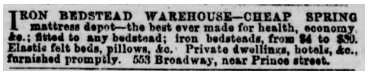

IRON BEDSTEAD WAREHOUSE—CHEAP SPRING mattress depot—the best ever made for health, economy &c.; fitted to any bedstead; iron bedsteads, from $4 to $80. Elastic felt beds, pillows, &c. Private dwellings, hotels, &c., furnished promptly. 553 Broadway, near Prince street.

Figure 6.8: Excerpt from the 9th October 1853 issue of the *New York Herald* https://chroniclingamerica.loc.gov/

The use of iron is not reserved only to industry. Across the entire corpus of the *New York Herald*, there are several daily notices of households looking for women to perform different household chores, including ironing. Additionally, a lot of furniture was made of iron. Figure 6.8 shows an advertisement from an 1853 copy of the *New York Herald* advertising for bedsteads made of iron. Such advertisements are common across all newspapers, selling a range of household items such as "a piano, built in iron, appropriate for beginners" (*Neue Hamburger Zeitung* 1914-06-12) or "iron stoves for inns or private houses" (*El Imparcial* 1883-01-22).

6.2.7 Conclusion

Iron is a material that was present in all aspects of Western society during the Second Industrial Revolution. Firstly, it was the material of the railway, which was both a method of transportation for a country's inhabitants and a huge investment opportunity for business people. During wartime, it was an essential material for creating weapons. But it was also the material of some of the most common and mundane household items such as kitchen utensils and beds. This causes conversation around the material to be very diverse and essentially neutral.

6.3 Telephone

The telephone was first patented by Alexander Bell in 1876 [27], which puts its invention during the Second Industrial Revolution. We can therefore find articles in several newspapers reporting about it for the first time, and follow its incorporation into society over time.

6.3.1 Beginnings


Tous les **TÉLÉPHONES** véritables du Professeur ALEXANDRE GRAHAM BELL, devront porter la marque de fabrique ci-après.

TÉLÉPHONE BELL

Les Téléphones non munis de cette marque sont des contrefaçons ou imitations et tous importateurs, fabricants et vendeurs seront poursuivis suivant la loi.

CORNELIUS ROOSEVELT,
seul Concessionnaire des Brevets pour la France.
1, Rue de la Bourse, PARIS.

Figure 6.9: Excerpt from a March 1878 issue of the *Le Figaro* https://gallica.bnf.fr/

When the telephone was first invented it caused great excitement and marvel. On the 20th of March 1977, *El Imparcial* reports to its readers that a test telephone call between Boston and Salem was successful and resulted in great applause. On the 19th of November 1877, a journalist from *Le Figaro* reports of his first experience with "the marvelous mechanism of the American discovery called telephone", which he describes as "fear-inducing but magnificent". Likely due to its extreme usefulness, telephones were soon installed in public spaces such as offices and administration. By March of 1878, you could find advertisements for telephones in *Le Figaro* (see Figure 6.9). In 1879, the *New York Herald* starts reporting of offices that have installed telephones "for the use and convenience of their patrons" (1879-08-03). In France, we can find telephone companies in stock market listings by the early 1880's.

6.3.2 Telephonic Demonstrations


TELEPHONE

will take place on MONDAY EVENING, April 2. Musical melodies will be performed in Philadelphia and distinctly heard by the audience in New York.
The following melodies will be performed Monday evening, April 2, on the Telephone:—
1. "Home, Sweet Home."
2. "Come Gentil" (Don Pasquale).
3. "Then You'll Remember Me" (Bohemian Girl).
4. "Last Rose of Summer."
5. "M'appari," romance (Martha).
6. Brindisi—"Il Segreto" (Lucrezia Borgia).

Figure 6.10: Excerpt from the 25th March 1877 issue of the *New York Herald* https://chroniclingamerica.loc.gov/

The telephone was such an innovation when it was first invented that going to see how it worked was a form of entertainment. The advertisement in Figure 6.10 from a 1877 issue of the *New York Herald* informs of a concert played in Philadelphia that will be transmitted to New York via telephone. In January 1878, *Le Figaro* writes of a theater performance that included a telephone conference in which the public could try out the news instrument (21-01-1878). Interestingly, the article mentions that many spectators were unconvinced of the utility and correct functioning of the device. An "Exposition of Experiences" hosted in the Champ-de-Mars in 1978 also allows visitors to try to telephone to Versailles, proving that "the problem of transmitting voice through large spaces is resolved" (*Le Figaro*, 1878-08-10).

6.3.3 Criticism

Especially in *Le Figaro*, there was a lot of criticism of the telephone in its first decades of existence and use. In 1881, an article complains that "it happens too often that the broken telephones are not repaired or replaced or that nobody responds to calls". The journalist did, however, call the device a "marvelous invention" (1881-06-18). Similarly, in 1893, an article states that "the subscription to the telephone is becoming unnecesserily expensive (1893-07-21). This criticism was not however completely limited to France. In 1899, the *Neue Hamburger Zeitung* stated that there was concern that public telephone booths could "easily spread disease", to which a group of experts responded that there were no such

risks if the devices were cleaned. Overall, it seems that there were quite a few reservations and issues with the telephone as it was first deployed.

6.3.4 Real Estate

When it was first invented, the telephone was much more prominent in public spaces than in private homes. The *Neue Hamburger Zeitung* writes in 1888 about an arrogant actor who "afforded himself the luxury of installing a telephone in his home" (1888-09-02), testifying to its unusual aspect. In *Le Figaro*, real estate listing start to include telephones in the late 1890's to early 1900's, although it was still rare until the 1910's. Even then, the houses that included telephones tended to be very large and expensive ones meant for the upper class ; the example in Figure 6.11 shows an advertisement for a waterfront villa on a 2-hectare property.


CABOURG à louer, meublée, 7,000 fr. saison, la belle villa « l'Oasis », confort mod., parc 2 hect., potager, téléphone, 1,500 mèt. casino, ligne tramway arrêt. Panoramas. S'adr. sur place et téléphone 130, à Neuilly-sur-Seine.

Figure 6.11: Excerpt from the 15th May 1918 issue of the *Le Figaro*
https://gallica.bnf.fr/

6.3.5 Commerce

Telephones obviously had a huge potential for commerce, allowing users to call companies to make reservations or inquire about products. Indeed, the device was very quickly used in this manner. In *Le Figaro*, advertisements start to have phone numbers, either for products or for performances, in the late 1890's, only 20 years after the telephone's invention. In *El Imparcial*, after the initial wave of wonder over the innovation had died down, the vast majority of mentions of the keyword are in relation to commercial advertisements, telling readers where to call to obtain a particular service or buy a particular object. In the *Neue Hamburger Zeitung*, the emergence of telephone numbers in advertisements happened in the early 1900's. Overall, by the 20th century, it seemed that all cities in which these newspapers are based were adapted enough to the new invention that it was normal to find telephone numbers in different types of listings.

A business that benefited particularly from this invention was the hotel business, especially in France. In *Le Figaro*, hotels listed that they had a telephones by the 1890s, as seen in Figure 6.12.

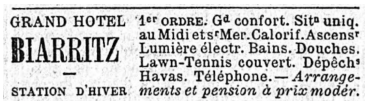

GRAND HOTEL BIARRITZ — STATION D'HIVER 1er ORDRE. Gd confort. Sita uniq. au Midi et sr Mer. Calorif. Ascenst Lumière électr. Bains. Douches. Lawn-Tennis couvert. Dépêchs Havas. Téléphone. — *Arrangements et pension à prix modér.*

Figure 6.12: Excerpt from the 1st December 1898 issue of the *Le Figaro*
https://gallica.bnf.fr/

6.3.6 Later years

Since the 1920's, the use of the word "telephone" had become more and more mundane in all newspapers. It seems that by then, all novelty of the device had worn off. The word was still mentioned in a large number of articles, but only to say that a certain piece of information was obtained or relayed by telephone. It was also more and more common to find telephone numbers as contact information in the pages listing real estate, job opportunities and object sales. We can say that by then, the telephone had definitely established itself in society.

Only in *Le Figaro* could we find a deviation from this trend. In the 1920's, the automatic telephone, which would not have to go through telephone centres, was brought to Paris. An article from the 23rd August 1922 reported that "certain post offices [...] were provided with special booths that contained automatic telephones". However, these were not

an instant hit. The same article states that "the public continues to push their constant complaints, no one is satisfied". Nevertheless, this attitude seemed to change in the following decade : an article from 1934 joyously reports "Telephone ! Telephones ! The automatic telephone is soon going to work in all of Paris" (1934-02-01).

6.3.7 Conclusion

This keyword was particularly interesting because we could track its beginnings in our corpus. It seems that all countries discover the telephone at the same time, the year after it was patented. Despite having its critics, it was clearly cause of marvel and excitement. It was used not only to communicate, but also to transmit concerts across cities. However, since it seemed to be adopted rather quickly in public administrations and offices, it quickly lost its novelty and was only mentioned in passing and neutrally later on.

Conclusion 7

7.1 Limitations

Since the project uses computation to analyse large amounts of digitised data, there are some limitations to the methods used.

Firstly, the data was digitised using OCR, which does not yield a perfect result. Indeed, especially in the case of the *Neue Hamburger Zeitung*, the quality of the OCR strongly limits what can be achieved with the data. The newspaper articles are also incomplete : they do not span the entire time frame evenly. This leads to gaps in our analysis for certain newspapers.

Secondly, most of the newspapers did not digitise their data article-by-article, but rather page-by-page or even issue-by-issue. An experimental pipeline was created to try and split these documents correctly, but it is far from perfect. Since different articles touch vastly different topics, not separating them correctly can confuse the topic modelling later-on.

Finally, all computational methods used are probabilistic, and as such prone to error. They may not take into consideration finer subtleties of language and semantics, which can be crucial to understanding an author's true point of view. They also assume that the given input text is fully correct, which is not the case given the errors induced by the OCR.

7.2 Conclusion and Outlook

The goal of this project was to analyse how the public reacted to different technologies during the Second Industrial Revolution across different countries. Many of the technologies invented, developed or expanded during this time like the telephone, the railway, electricity, cars and many more had a huge impact on society. It is therefore interesting to explore not only what the public thought of these innovations when they were still new, but also to compare these opinions across different countries to get an overall idea of public opinion in these matters in late 19th and early 20th century Western society.

This analysis was based on a corpus of newspapers from France, Germany, Spain and the United States. From this corpus were extracted articles referencing certain technologies, namely *electric light*, *gasoline*, *iron* and *telephone*. These articles were then split into several time-spans in order to compare the topics discussed in them through time and across countries. Since there were only very few articles discussing electric light, this keyword was discarded from further analysis.

To detect topics in these articles, a separate Pachinko Allocation Model was created for each combination of newspaper, keyword and time-span. These were projected onto a multi-dimensional space representing

semantic closeness using word embeddings. To compare these word embeddings across languages, MUSE multi-lingual word embeddings were used. Semantic "distances" between different topics, represented as a set of words, could then be measured across newspapers and time-spans using Word Mover's Distance. These distances could then be interpreted as a network, which in turn could be translated into topic communities for every keyword.

The analysis shows that *gasoline* was most commonly discussed in the context of the stock market. In almost every country, one can find the term in the long lists describing current market values that were featured in these newspapers. However, since the material was later a key component for cars and airplanes, we can also find discussions of those technologies in the later issues of our newspapers. They were mostly related to excitement : cars were associated with speed and racing, and airplanes were a fascinating novelty.

Iron was a particularly important material during the Second Industrial Revolution because it was the material of the railway. Being a time of huge expansions in this area, the word was discussed in both an economical context and in advertisements letting readers know the local train schedule. It was also used in real estate and furniture, meaning that practically any person living in this time was in close contact with the material.

The case of the *telephone* is particularly interesting. It was invented within our analysed time span and as such, we can track its evolution from a wonderful novelty to a normalised everyday tool. At first, it was so exciting to the population that it was considered a form of entertainment in itself. However, it didn't lack critics, especially in France. By the 1920's, the novelty of the device had worn off, and whilst it was still mentioned constantly, it was only in passing and not with the same tone of marvel that was used when it was first brought to the public.

This project aimed to analyse the evolution, similarities and deviations of public opinion on specific technologies across time and countries. It could be further examined by better splitting the newspaper articles, or by performing sentiment analysis to get a more specific reading of the tone used in these discussions.

Bibliography

Here are the references in citation order.

[1] Jürgen Habermas, Sara Lennox, and Frank Lennox. 'The Public Sphere: An Encyclopedia Article (1964)'. In: *New German Critique* 3 (1974), pp. 49–55. (Visited on 06/09/2023) (cited on page 1).

[2] Joel Mokyr and Robert H Strotz. 'The second industrial revolution, 1870-1914'. In: *Storia dell'economia Mondiale* 21945.1 (1998) (cited on page 1).

[3] Eric Hobsbawm. *The Age of Revolution 1789-1848*. 1st Vintage Books ed. Vintage Books, 1996 (cited on pages 2, 15).

[4] Eric Hobsbawm. *The Age of Capital 1848-1875*. Weidenfeld and Nicolson, 1975 (cited on page 2).

[5] The Editors of Encyclopaedia. Britannica. 'Le Figaro'. In: *Encyclopedia Britannica* (2017) (cited on page 3).

[6] Julio Rodríguez Puértolas Carlos Blanco Aguinaga and Iris M Zavala. *Historia social de la literatura española*. Vol. 56. Ediciones Akal, 2000 (cited on page 3).

[7] C. Sonntag. *Medienkarrieren: biografische Studien über Hamburger Nachkriegsjournalisten 1946-1949*. Forum Kommunikation und Medien. M Press, 2006 (cited on page 3).

[8] James L. Crouthamel. *Bennett's New York Herald and the Rise of the Popular Press*. Syracuse University Press, 1989. (Visited on 05/31/2023) (cited on page 3).

[9] Thomas Bench. *anxietyNews*. https://github.com/ThomasBench/anxietyNews. 2022 (cited on page 4).

[10] Elisa Michelet. 'An Industrial West? Analyzing Multilingual Newspapers Discourses about Technology during the Second Industrial Revolution (1840-1930)'. In: *EPFL* (2015) (cited on pages 5, 7–11).

[11] Rubayyi Alghamdi and Khalid Alfalqi. 'A survey of topic modeling in text mining'. In: *Int. J. Adv. Comput. Sci. Appl.(IJACSA)* 6.1 (2015) (cited on page 7).

[12] Germans Savcisens. *Simple Topic Modelling Examples*. https://github.com/carlomarxdk/topic_modelling. 2022 (cited on pages 7, 8).

[13] David M Blei, Andrew Y Ng, and Michael I Jordan. 'Latent dirichlet allocation'. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022 (cited on page 7).

[14] Wei Li and Andrew McCallum. 'Pachinko allocation: DAG-structured mixture models of topic correlations'. In: *Proceedings of the 23rd international conference on Machine learning*. 2006, pp. 577–584 (cited on page 7).

[15] Wei Li, David Blei, and Andrew McCallum. 'Nonparametric bayes pachinko allocation'. In: *arXiv preprint arXiv:1206.5270* (2012) (cited on page 7).

[16] David Newman et al. 'Automatic evaluation of topic coherence'. In: *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*. 2010, pp. 100–108 (cited on page 7).

[17] Eva Cantoni et al. 'Longitudinal variable selection by cross-validation in the case of many covariates'. In: *Statistics in medicine* 26.4 (2007), pp. 919–930 (cited on page 8).

[18] Matthew Honnibal and Ines Montani. 'spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing'. To appear. 2017 (cited on page 8).

[19] bab2min. *Tomotopy*. https://github.com/bab2min/tomotopy. 2019 (cited on page 8).

[20] Dhruvil Karani. 'Introduction to word embedding and word2vec'. In: *Data Sci* 1 (2018) (cited on page 9).

[21] Piotr Bojanowski et al. 'Enriching Word Vectors with Subword Information'. In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146 (cited on page 9).

[22] Tomas Mikolov et al. 'Efficient estimation of word representations in vector space'. In: *arXiv preprint arXiv:1301.3781* (2013) (cited on page 9).

[23] Alexis Conneau et al. 'Word Translation Without Parallel Data'. In: *arXiv preprint arXiv:1710.04087* (2017) (cited on page 9).

[24] Matt Kusner et al. 'From Word Embeddings To Document Distances'. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 957–966 (cited on page 10).

[25] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. 'Finding community structure in very large networks'. In: *Physical review E* 70.6 (2004), p. 066111 (cited on page 11).

[26] Pasquale De Meo et al. 'Generalized louvain method for community detection in large networks'. In: *2011 11th international conference on intelligent systems design and applications*. IEEE. 2011, pp. 88–93 (cited on page 11).

[27] Library of Congress Science Reference Section. *Who is credited with inventing the telephone?* https://www.loc.gov/everyday-mysteries/technology/item/who-is-credited-with-inventing-the-telephone. 2022 (cited on page 18).